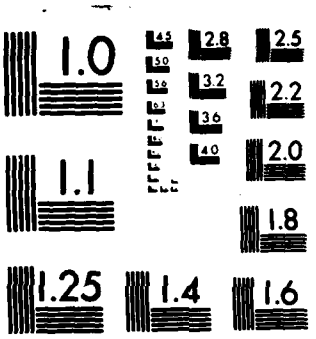
AD-A129 599 OPTIMUM QUANTIZATION OF FIR WIENER AND MATCHED FILTERS 1/1
(U) MOORE SCHOOL OF ELECTRICAL ENGINEERING PHILADELPHIA
PA DEPT O.. C CHEN ET AL. 1983 AFOSR-TR-83-0506
UNCLASSIFIED AFOSR-82-0022 F/G 12/1 NL



UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(2)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFOSR-TR- 83-0506	2. GOVT ACCESSION NO.: AD-A129 599
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): OPTIMUM QUANTIZATION OF FIR WIENER AND MATCHED FILTERS	5. TYPE OF REPORT & PERIOD COVERED: TECHNICAL
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): C.T. Chen and S.A. Kassam	8. CONTRACT OR GRANT NUMBER(s): AFOSR-82-0022
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Systems Engineering, University of Pennsylvania, Philadelphia PA 19104	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: PE61102F; 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS: Directorate of Mathematical & Information Sciences, Air Force Office of Scientific Research, Bolling AFB DC 20332	12. REPORT DATE: June 1983
	13. NUMBER OF PAGES: 4
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited.

DTIC ELECTE JUN 22 1983 S D

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

B

18. SUPPLEMENTARY NOTES
Proc. IEEE Inter. Conf. on Communications, June 1983. p. 1-4

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Optimum quantization; FIR filters; Wiener filters; matched filters.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
Quantization schemes are considered for the coefficients of discrete-time finite-impulse-response filters for estimation and detection. The quantized filters are optimized with respect to estimation or detection performance criteria, and recursive algorithms are developed for use in finding numerical solutions. Results indicate that in general only a few levels of quantization can give very good performance.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AD A129599

DTIC FILE COPY

OPTIMUM QUANTIZATION OF FIR WIENER AND MATCHED FILTERS*

Cheng-Tie Chen and Saleem A. Kassam

The Moore School of Electrical Engineering
University of Pennsylvania
Department of Systems Engineering
Philadelphia, Pennsylvania 19104

ABSTRACT

In this paper quantization schemes are considered for the coefficients of discrete-time finite-impulse-response filters for estimation and detection. For filter impulse-response sequences of length J, we consider the optimum choice of a smaller number K of coefficient values and their distribution over the J-sample filter sequence. The quantized filters are optimized with respect to estimation or detection performance criteria, and recursive algorithms are developed for use in finding numerical solutions. Results we give indicate that in general only a few levels give good performance.

I. INTRODUCTION

The use of matched filters and Wiener filters is widespread as signal processing elements in many applications. In many situations such filtering is performed on discrete-time data sequences, and the filter is implemented as the convolution of a finite impulse-response sequence with the data sequence. Wiener filtering is performed when the data represents a noisy random signal which is to be estimated with minimum mean-square-error(MSE), whereas matched filtering can be used to maximize the output signal-to-noise ratio (SNR) at specific times when the input is a noisy version of some deterministic signal.

Let $\{x_t\}$ represent an observation sequence of a signal $\{v_t\}$ and additive zero-mean noise $\{w_t\}$. Suppose this sequence is to be convolved with a finite impulse-response sequence $\{g_t\}_0^{J-1}$ of length J. In estimation problems the output $\{y_t\}$ may be required to be an estimate of $\{v_{t-n}\}$ for some integer n. Similarly in matched filtering, the output values $\{y_{(J+m)t}\}$ for some fixed repetition interval J+m may be required to have maximum SNR values. In general, optimization of the filter impulse-response sequence results in a specific sequence of J different numerical values. In many applications, however, we may be interested in using a smaller number K of distinct values in forming the filter impulse-response sequence of length J. This may be desirable (a) to allow the filter to be easily updated in adaptive systems, (b) as a first step in obtaining sub-optimum but computationally efficient schemes, or (c) to provide a degree of robustness of performance under deviations from assumed signal and noise characteristics.

*This research is supported by the Air Force Office of Scientific Research under Grant AFOSR 82-0022.

We will refer to partitioning of the J impulse-response samples into K groups, with each of which a distinct level is associated, as quantization of the impulse-response sequence. In considering optimum quantization it is natural to use as a criterion of performance the original criterion of minimizing the MSE or maximizing SNR. In addition one can also look for a "best-fit" quantization of the optimum filter, for example by minimizing mean-square-deviation between the optimum and quantized impulse-response sequences. We will consider these quantization schemes and give numerical performance results in the rest of this paper. It will be seen that in many cases system performance with a low-order quantization can be expected to be quite close to optimum performance.

In earlier work [1] quantization of the coefficients of a matched filter was considered, for white noise and low input SNR. In [2] quantization of the Wiener filter frequency response was considered. In another recent paper [3] computationally efficient "multiplication-free" implementations of quantized Wiener filters and equalizers have been considered.

II. OPTIMUM QUANTIZED FIR WIENER FILTERS

With our previous notation, the output of an FIR filter is

$$y_t = \sum_{i=0}^{J-1} x_{t-i} g_i .$$

Defining $h_i^* = g_{i-1}$ and $s_i = v_{t-i+1}$, $n_i = w_{t-i+1}$ and $r_i = x_{t-i+1}$ for i=1,2,...,J, we have $y_t = \underline{h}^\dagger \underline{r} = \underline{h}^\dagger(\underline{s} + \underline{n})$ where † means conjugate-transpose and where the vectors are column vectors of corresponding components indexed from 1 to J. For Wiener filtering (with zero-mean $\underline{s}$), let the quantity to be estimated at time t be $d = \underline{h}_d^\dagger \underline{s}$, where $\underline{h}_d$ is some given J-vector. For example, $\underline{h}_d$ could be [1,0,..., 0,0] to estimate v_t from $x_t, x_{t-1}, \ldots, x_{t-J+1}$. For a given filter $\underline{h}$, the MSE e(h) between y_t and d can be obtained as

$$e(\underline{h}) = \underline{h}_d^\dagger S \underline{h}_d + \underline{h}^\dagger R \underline{h} - 2\,\mathrm{Re}\{\underline{h}^\dagger C \underline{h}_d\}, \qquad (1)$$

Presented at the International Conference on Communication, June, 1983.
To be published in the IEEE Proceedings.

where $S = E\{\underline{s}\,\underline{s}^{\dagger}\}$, $R = E\{\underline{r}\,\underline{r}^{\dagger}\}$ and $C = \{\underline{r}\,\underline{s}^{\dagger}\}$. The filter $\underline{h}_0$ minimizing $e(\underline{h})$ is given by

$$\underline{h}_0 = R^{-1}\underline{Ch}_d.$$

For any K-th order quantized filter (with K groups) define a grouping matrix $Q = [Q_{kj}]_{KxJ}$ such that $Q_{kj} = 1$ if $j\epsilon I_k$, the k-th of K groups, and $Q_{kj} = 0$ otherwise. Each column of Q has a single non-zero entry. Let the level associated with the k-th group be H_k, k=1,2,...K. Thus $h_j = H_k$ if $Q_{kj} = 1$. From this it follows that for such a quantizated filter $\underline{h} = Q^{\dagger}\underline{H}$ where the level-vector $\underline{H} = [H_1, H_2, \ldots H_K]^T$.

For a K-th order quantized filter the MSE between y_t and d is obtained by replacing $\underline{h}$ with $Q^{\dagger}\underline{H}$ in (1). For given Q the optimum level-vector $\underline{H}_0$ is easily obtained to be

$$\underline{H}_0 = (QRQ^{\dagger})^{-1}QCh_d, \qquad (2)$$

and an expression for the resulting MSE can be found directly from (1).

Next we consider the case where the quantized filter levels are given. This is useful in practice when the levels are fixed for simplicity of implementation. The following theorem gives a necessary condition on the optimum grouping to minimize the MSE.

Theorem I. Let the level-vector $\underline{H}$ of K distinct levels be given, and let I_{ok}, k = 1,2,...,K denote the corresponding optimum groups. If $p\epsilon I_{om}$, then

$$\mathrm{Re}\{(H_n^* - H_m^*)[\sum_{k=1}^{K} 2H_k \sum_{\substack{q\neq p \\ q\epsilon I_{ok}}} R_{pq} + (H_n + H_m)R_{pp} - 2(Ch_d)_p]\} \geq 0, \qquad (3)$$

for all $n \neq m$, where R_{pq} is the pq-th element of R and $(Ch_d)_p$ is the p-th element of Ch_d. If the equality in (3) holds for a specific n, p can be removed from I_{om} and assigned to I_{on} without changing the MSE.

The proof can be found in [4]. When R is diagonal, the first term in the bracket in (3) vanishes, and (3) is essentially a necessary and sufficient condition [within the ambiguity caused by the equality in (3)]. From the above, optimum quantized filter levels and optimum groups which jointly minimize the MSE must satisfy (2) and (3) simultaneously. Note that (2) and (3) provide only necessary conditions for the optimum quantized Wiener filter.

The results in Theorem I can be simplified if the quantities involved are real. This is given by the following theorem.

Theorem II. Assume that all quantities are real, and assume an indexing for the given filter levels such that $H_m < H_{m+1}$, m=1,2,...,K-1, without loss of generality. Let I_{om} be the corresponding m-th optimum group. If $p\epsilon I_{om}$, then

$$b_{m-1} \leq f(p) \leq b_m \qquad (4)$$

where $b_m = (H_m + H_{m+1})/2$ for m=1,2,...,K-1 ($b_0 = -\infty$, $b_K = \infty$), and

$$f(p) = \{(Ch_d)_p - \sum_{k=1}^{K} H_k \sum_{\substack{q\neq p \\ q\epsilon I_{om}}} R_{pq}\}/R_{pp}. \qquad (5)$$

If the lower(upper) equality in (4) holds, p can be removed from I_{om} and assigned to $I_{o(m-1)}$ ($I_{o(m+1)}$) without increasing the MSE.

The proof is given in [4]. Now it is obvious that for the case of real quantities, the optimum quantized levels H_{oi} (indexed in increasing order) and the corresponding optimum groups I_{oi} must satisfy simultaneously (2) and (4). Based on these necessary conditions, a recursive algorithm can be developed to find particular solutions for H_{oi} and I_{oi}.

Algorithm I.

1) Initialize: Q^1 = KxJ null matrix.
   Input: Initial guess of grouping matrix, Q.
2) Find $\underline{H}_o$ from (2). Re-index to get elements of $\underline{H}_o$ in rank order. Interchange rows of Q likewise.
3) If $Q = Q^1$, stop. Current $\underline{H}_o$ and Q then represent candidate for optimum quantization. Otherwise, set $Q^1 = Q$ and continue.
4) Compute $\underline{b}$ vector and f(p), p=1,2,...,J of Theorem 2. Assign p to m-th group if $b_{m-1} < f(p) \leq b_m$, and obtain new Q. Go to step 2.

Different initial guesses of Q may lead to different candidates for the optimum quantization, since the algorithm is based on the necessary condition for optimality. In practice, from several initial guesses one can pick the best result obtained. The algorithm generally converges rather quickly when it does converge, and can be used to obtain at least very good sub-optimum schemes when the MSE's of the non-quantized and resulting quantized filters are close. Note that minor modifications in the above algorithm allow it to obtain candidates for the optimum Q for given $\underline{H}$.

The algorithm will fail to converge if the result at the r-th step is the same as the result at the s-th step for some $r > s+1$. If at some step one or more groups contain no element-index, i.e. if one or more rows of Q are all zero, the iterations cannot continue because then $QRQ^{\dagger}$ cannot be inverted. This can happen because of a bad choice for the initial guess or if the optimum (unquantized) filter requires less than K distinct levels.

III. OPTIMUM QUANTIZED FIR MATCHED FILTERS

With the same notation as that used in Section II, we find that the SNR for the output y_t is

$$\rho(\underline{h}) = |\underline{h}^{\dagger}\underline{s}|^2/\underline{h}^{\dagger}N\underline{h} \qquad (6)$$

where $N = E\{\underline{n}\,\underline{n}^{\dagger}\}$. The filter $\underline{h}_o$ maximizing $\rho(\underline{h})$ is $\underline{h}_o = N^{-1}\underline{s}$. For a K-th order quantized filter

the output SNR is obtained from (6) by replacing $\underline{h}$ with $Q^{\dagger}\underline{H}$. For given Q the optimum level-vector $\underline{H}_o$ is easily obtained to be

$$\underline{H}_o = (QNQ^{\dagger})^{-1} Q\underline{s}, \tag{7}$$

and an expression for the resulting SNR can be found directly from (6).

Unlike the Wiener filtering problem, no explicit result on the optimum grouping has been obtained for the case of given quantized filter levels. However, a necessary condition similar to that in the previous section can be found if we consider jointly the optimum quantized filter levels and the optimum grouping.

Theorem III. Let H_{oi}, I_{oi}, i=1,2,...,K denote the optimum quantized filter levels and the optimum groups which jointly maximize the output SNR. If $p\epsilon I_{om}$, then H_{oi} and I_{oi} satisfy (7) and the following inequality:

$$\mathrm{Re}\{(H_{on}^{*}-H_{om}^{*})[\sum_{k=1}^{K} 2H_{ok}\sum_{q\neq p,\ q\epsilon I_{ok}} N_{pq} + (H_{on}+H_{om})N_{pp}-2s_p]\} \geq 0 \tag{8}$$

for all $n \neq m$. If the quantities involved are real, we can reduce (8) (by assuming $H_{oi} < H_{o(i+1)}$ without loss of generality) into

$$b_{m-1} \leq f(p) \leq b_m \tag{9}$$

where $b_m = (H_{om} + H_{o(m+1)})/2$ for m=1,2,...,K-1 ($b_o = -\infty$, $b_K = \infty$), and

$$f(p) = \{s_p - \sum_{k=1}^{K} H_{ok}\sum_{q\neq p,\ q\epsilon I_{ok}} N_{pq}\}/N_{pp}. \tag{10}$$

The proof is given in [4]. It is interesting to note the similarities between (8) and (3), or (9) and (4). Hence, for the case of real quantities, an algorithm similar to Algorithm I can be developed to find specific candidates for the optimum quantized matched filter.

IV. BEST-FIT QUANTIZATION OF OPTIMUM FILTERS

One approach to finding a reasonable quantization of a J-component optimum filter $\underline{h}_o$ into a K-valued filter is by seeking that K-th order filter which minimizes some measure of "distance" between $\underline{h}_o$ and its quantized version. In particular, consider the integrated-squared-error (ISE) measure

$$e_s = \sum_{k=1}^{K}\sum_{q\epsilon I_k} |H_k - h_{oq}|^2. \tag{11}$$

For given grouping, i.e. I_k, k=1,2,...,K, it is easy to show that the minimum ISE is obtained for levels H_{ok} satisfying

$$H_{ok} = \sum_{q\epsilon I_k} h_{oq} / \sum_{q\epsilon I_k} 1. \tag{12}$$

Thus H_{ok} is the average of the optimum levels for the k-th group. For given level-vector $\underline{H}$ we have the following:

Theorem IV. Let I_{oi}, i = 1,2,...,K, denote the optimum groups minimizing e_s for given quantized filter level-vector $\underline{H}$. If $p\epsilon I_{om}$, then

$$\mathrm{Re}\{(H_n^{*} - H_m^{*})h_{op}\} \leq 1/2\ (|H_n|^2 - |H_m|^2) \tag{13}$$

for all $n \neq m$. In addition, if (13) is true with strict inequality for all $n \neq m$ and some p, then $p\epsilon I_{om}$. If, for some p, (13) is true for all $n \neq m$ but the equality in (20) holds for some specific n=j, then p can belong to either I_{om} or I_{oj}. Moreover, if the quantities involved are real and $H_1 < H_2 < \ldots < H_K$, then (13) reduces to

$$b_{m-1} \leq h_{op} \leq b_m \tag{14}$$

where $b_m = (H_m + H_{m+1})/2$ for m=1,2,...,K-1 ($b_o = -\infty$, $b_K = \infty$).

The results in Theorem IV are obtained from the property that $p\epsilon I_{om}$ if and only if the summand $|H_m - h_{op}|^2$ in (11) is the smallest among all $|H_k - h_{op}|^2$, k=1,2,...K. The proof is omitted.

From the above, the best-fit quantized filter levels H_{oi} and the best-fit groups I_{oi} which jointly minimize the ISE must satisfy (12) and (14) simultaneously for the case of real quantities. These two equations provide only necessary conditions. An algorithm similar to Algorithm I can be developed to find particular candidates for the best-fit quantized filter.

Before we proceed to give numerical examples, we note that the best-fit quantization of optimum filters is not, in general, the same as the optimum quantization discussed in Sections II and III. Clearly, the best-fit quantized filter will have somewhat higher MSE in Wiener filtering, and somewhat lower SNR in matched filtering.

V. NUMERICAL EXAMPLES

The two examples in this section illustrate the use of the results we have obtained.

Example 1 (Wiener Filter)

In this example we consider a 15-sample estimation problem with $\underline{h}_d = [1,0,0,\ldots,0]^T$. The signal and noise are uncorrelated so that R=S+N and C=S, where the ij-th elements of the signal and noise covariance matrices S and N are $8.0\exp(-0.3\cdot|i-j|)$ and $6.0\exp(-0.8|i-j|)$, respectively. The optimum Wiener filter coefficient vector is $\underline{h}_o = [0.497,\ 0.729\times10^{-1},\ 0.434\times10^{-1},\ 0.259\times10^{-1},\ 0.154\times10^{-1},\ 0.919\times10^{-2},\ 0.548\times10^{-2},\ 0.326\times10^{-2},\ 0.195\times10^{-2},\ 0.116\times10^{-2},\ 0.691\times10^{-3},\ 0.413\times10^{-3},\ 0.247\times10^{-3},\ 0.149\times10^{-3},\ 0.137\times10^{-3}]^T$, giving an MSE of 3.251.

For second-order quantization (K=2) different initial guesses resulted in basically two types of quantization schemes with good MSE performance. The

best MSE of 3.364 was obtained with filter levels of (1) 0.942×10^{-2} and (2) 0.553, distributed in the obvious way over the 15-sample impulse response sequence as {2 1 1 1 ... 1}. Many of the runs of Algorithm I with different initial guesses gave a quantizer with levels (1) 0.362×10^{-2} and (2) 0.304, distributed as {2 2 1 1 1 ... 1}, with an MSE of 3.705. For K=3 basically very similar results were obtained, with improvements in MSE performance which are not major because of the good results for K=2.

For this example one conclusion from the above is that using only the current sample with a weight of 0.5 one can expect good results. In fact, the resulting MSE is 3.5. However, the results above also indicate that good performance may be obtained by using only levels of 0.3 for the first two samples (current and previous sample), with zero weighting for the others. The resulting MSE is 3.72. The possibility of such a scheme is not obvious without the algorithm. This type of scheme may be preferable if the data is subject to occassional random higher-variance contamination, because the averaging over two samples will then provide a better estimate. In general the use of Algorithm I allows generation of a set of good quantizer designs from which a specific useful or easily implemented approximation may be derived. For simple implementation the modification of Algorithm I may be used to determine optimal grouping once a good set of levels is determined.

For this example with K=2 the best-fit quantizer, as given by the necessary conditions, turned out to have levels of (1) 0.129×10^{-1} and (2) 0.497 for all the different initial guesses that were tried. The distribution over the 15 samples was the obvious one, and the resulting MSE was 3.407. While this was very close to one of two the previous results, note that the second quantizer would not have emerged from the best-fit criterion.

Example 2 (Matched Filter)

Again we use J=15. The deterministic signal vector $\underline{s}$ is formed from uniformly spaced samples of an amplitude-tapered sinusoidal waveform. Specifically, we take $s = 2.5 \cos[0.2\pi(i-1)] \cdot \cos[0.025\pi |i-1|]$. The ij-th element of the input noise covariance matrix is assumed to be $0.5 \exp(-0.8|i-j|)$. For this case the optimum matched filter coefficient vector is $h = [3.995, 2.399, 0.874, -0.956, -2.347, -2.761, -2.085, -0.659, 0.892, 1.938, 2.114, 1.454, 0.343, -0.693, -1.846]^{t}$, giving the maximum SNR of 42.61.

For third-order quantization (K=3) different initial guesses for Q in the recursive algorithm again gave several different results which were very good in SNR performance. In all cases convergence took place in less than 7 iterations. However, basically two distinct types of quantized filters emerge from the different initial guesses. Of the twenty different trials made, the best SNR obtained for K=3 was 40.11, obtained with levels (1) -1.791, (2) 1.178, and (3) 3.003, distributed over the 15-sample impulse-response as {3 3 2 1 1 1 1 1 2 2 3 2 2 1 1}. Many initial guesses gave optimum levels of (1) -1.777, (2) 0.960, and (3) 2.699, distributed as {3 3 2 1 1 1 1 1 2 3 3 2 2 1 1} with an SNR of 39.72. Another basically similar type of quantizer (with levels close to -2.0, 1.0 and 3.0) gave an SNR of 39.96.

Another group of initial guesses gave, for K=3, the levels (1) -2.187, (2) -0.172, and (3) 2.293, distributed as {3 3 3 1 1 1 1 2 2 3 3 3 2 2 1}, with an SNR of 39.54. Again, several other guesses gave this type of result (with levels close to -2.0, 0 and 2.0) with SNR's very close to 39.54. One interesting indication from this is that the use of levels -2, 0, 2 distributed in this way should give good results, and would require only one bit signed coefficients. Indeed, the SNR for this scheme turns out to be 39.46. Thus one use for the theoretical results is for providing a set of good quantization schemes on which a choice of a simple scheme may be based.

The candidates for best-fit minimum ISE quantizers were also obtained for this example. Interestingly, while the specific numerical values are different, it turns out that the resulting quantizers also fall into one of the same two basic types. The best quantizer by this criterion, from the initial guesses tried, gives levels of (1) -1.621, (2) 0.891 and (3) 2.611, distributed as for the similar quantizer based on maximizing SNR, and gives an SNR of 39.69. Thus we may conclude that the best-fit quantizer gives very good results for such examples. This implies that it should be reasonable to find optimum grouping for given levels using the minimum ISE criterion, since no such analytical result has been found for maximizing SNR.

REFERENCES

1. S.A. Kassam and T.L. Lim, "Coefficient and Data Quantization in Matched Filters for Detection," IEEE Trans. Communication, Vol. COM-26, pp. 124-127, January 1978.
2. L.J. Cimini and S.A. Kassam, "Optimum Piecewise Constant Wiener Filters," J. Opt. Soc. Am, Vol. 71, pp. 1162-1171, October 1981.
3. T.Y. Yan and K. Yao, "A Multiplication-Free Solution for Linear Minimum Mean-Square Estimation and Equalization Using the Branch-and-Bound Principle," IEEE Trans. Information Theory, Vol. IT-26, pp. 316-326, May 1980.
4. C.-T. Chen, "Robust and Quantized Linear Filtering for Multiple-Input Systems," Ph.D. dissertation, Dept. of Systems Engineering, Moore School of Electrical Engineering, Univ. of Pennsylvania, 1983.